

# Examiners' Report Principal Examiner Feedback

Summer 2023

Pearson Edexcel GCE In Statistics (9ST0) Paper 02: Statistical Inference

# **Edexcel and BTEC Qualifications**

Edexcel and BTEC qualifications are awarded by Pearson, the UK's largest awarding body. We provide a wide range of qualifications including academic, vocational, occupational and specific programmes for employers. For further information visit our qualifications websites at <u>www.edexcel.com</u> or <u>www.btec.co.uk</u>. Alternatively, you can get in touch with us using the details on our contact us page at <u>www.edexcel.com/contactus</u>.

# Pearson: helping people progress, everywhere

Pearson aspires to be the world's leading learning company. Our aim is to help everyone progress in their lives through education. We believe in every kind of learning, for all kinds of people, wherever they are in the world. We've been involved in education for over 150 years, and by working across 70 countries, in 100 languages, we have built an international reputation for our commitment to high standards and raising achievement through innovation in education. Find out more about how we can help you and your students at: www.pearson.com/uk

Summer 2023 Publications Code 9ST0\_02\_2306\_ER All the material in this publication is copyright © Pearson Education Ltd 2023

## **General Comments**

This paper was the third full sitting of the new A level specification. Throughout the marking it was shown that the paper was accessible to all candidates and that the paper could be completed within the time allowed.

It is important to make candidates aware that working in the tables will be seen and marked by examiners. Therefore, they do not need to use time in the exam by rewriting the tables to then rank or calculate expected values etc.

It is important for students to be able to comment on reliability and validity to acquire the A03 marks in this specification. Although there were instances in the paper where candidates were successful, for instance identifying the use of Central Limit Theorem in Q03(c), it was apparent that this was a discriminating skill at the top end.

Candidates should continue to be reminded that explanation answers should be given in the context of the question.

## **Question 1**

This question proved to be a good settler for the exam, with many candidates achieving full marks for the Wilcoxon Rank Sum tests in (a). It was pleasing to see several candidates clearly identifying the hypothesis test in their answers, though this was not required in the mark scheme. The use of the wording 'non-parametric test' in the question was identified by many students and enabled them to correctly identify the required test. There were instances of candidates conducting a difference of two mean *t*-test, which could not achieve any marks in (a).

A common misconception across candidates was whether the value of 0 is included in the ranking of the samples. This is the third time that a 0 value has been given in the new specification and it is important that candidates are aware that the **0 is included within the ranking**. Those who did not include it lost accuracy marks.

Part (b) was answered very well. Credit was given for having equal sample sizes. However, it should be noted that equal sample sizes would only improve the test if the overall sample size is maintained, or increased. Some candidates suggested an alternative test, it is important that the wording of the question is noted 'an improvement for the test in (a)' therefore answers must be improvements to the Wilcoxon Rank Sum test.

Part (c) was attempted by the majority of candidates with many being able to identify the need for a **normally distributed population**, use of the words it, data or sample were not penalised if used instead of population at this stage in the paper. Credit was also given for candidates who mentioned the need for a **common variance for the populations**.

The final part of this question proved to be the most challenging aspect of the questions. Though it was pleasing to see many candidates picking up at least 1 mark in this part. It is important for candidates to be aware that their responses must be **specific and in context**. With regards to the living arrangements comment there were candidates who said 'this was not valid as the improvement was likely due to practice' this did not gain credit as they themselves have assumed a cause that was not tested in (a). Some candidates criticised the test rather than the comments that had been made.

## Question 2

Again, the contingency table method has proved to be accessible to many candidates. Though some students struggled with the hypothesis test being broken up across several pages of the paper.

Part (a) was attempted by a significant number of candidates and the majority scored these marks. It would be useful to remind candidates to **double check** calculations on these more accessible questions.

The completion of the hypothesis test in (b) was done well by many candidates using a variety or methods. For those who are using the calculator it is important that they **write down the test statistic contributors**. This is important so that method marks can be awarded if errors have been made in the calculator work, but also so that contributors can be identified for association questions which may follow, as in (c)

As mentioned in last years report the main errors of candidates not gaining full credit were for hypotheses being reversed (this was seen more frequently in candidates using the word independence rather than association) or selection of the incorrect critical value.

Part (c) was answered well by the top end of candidates. For association questions it is required for the candidate to identify **the largest test statistic contributor**, those who scored well on this part identified the test statistic as well as the relevant observed and expected values.

Candidates well practised in exam technique were able to pick up on the key word of '**describe**' in the question. Therefore, those who simply identified Brisbane and Pink did not gain credit, there needed to be an attempt at saying that more females chose pink in Brisbane than would be expected.

Again, the most challenging part of the question was (d). The first mark required candidates to identify that the association found may now be due to age rather than location, bringing in the concept of gender stereotypes was a common incorrect answer unless the candidate directly explained the effect on the conclusion. The second mark was achieved more frequently, though it was important that candidates specifically mentioned dealing with the ages in their answer rather than other possible variables.

'Introduce a blocking factor of age' was a very common incorrect answer for (d)

#### **Question 3**

This question was particularly challenging, though it was pleasing to see that the vast majority of candidates had attempted all parts of the question.

The paper did not specifically state a particular test in (a). Therefore, this was left for interpretation by candidates. However, it is important that candidates are encouraged to select the **most powerful** test with **fewest assumptions** where appropriate, this was reflected in the marking. Those who attempted either a *t*-test or Wilcoxon signed rank test could achieve all 6

marks. Those who attempted a *z*-test achieved a maximum of 4 marks (2 assumptions made) and those who attempted the sign test achieved a maximum of 3 marks (no assumptions but least powerful).

For both (a) and (b) it is important that candidates using hypotheses like  $H_0$ :  $\mu_A = \mu_B$  that the subscripts are clearly identified in their method. Hypotheses like  $H_0$ :  $\mu_F = \mu_M$  gained full credit as it was identifiable by the context.

A significant number of candidates correctly identified that the large sample would lead to a z test in (b) and this proved to be completed more successfully than (a).

Both *z* and *t* tests are tests on an average (specifically the mean). Therefore, candidates should be referring to the mean in their conclusions.

It was pleasing to see many fully correct responses to (c). Those who only identified the use of central limit theorem but not how this was relevant to the assumptions achieved 1 mark.

In the final part of the question was assessing the candidate's ability to communicate findings effectively. Although it was challenging to gain full credit many candidates achieved 3 marks on this part. **Follow through marks** could be achieved for this part (even without completing the tests in (a) and (b)) it is therefore prudent to encourage candidates to use any information that they can to answer these parts (those who just compared the means directly were awarded credit for a correct explanation in context).

A common misconception was that 'Do not reject  $H_0$ ' meant that there were more goals in the premier league than in the super league.

The audience mark was penalised for use of: random sample, mean, variance, standard deviation, 5% significance or naming a specific test.

For the second mark candidates could either comment on the values seen in the sample or indicate that this result is for 2022 and has therefore changed since 2019.

#### **Question 4**

A new style of question that saw those candidates with a more refined exam technique scoring well. In questions where specific bullet points are given, in this case 3, the responses given must reference **all 3 bullet points**.

Those who scored well broke their answer up into: Similarities, Differences and Assumptions, this meant that comments were **clearly identified** as one of these 3 categories.

Some candidates wrote in long prose with a general description of the test. These did not gain credit as it is the candidate who is required to clearly identify similarities and differences, not the examiner.

Similarities and differences that could have been true for other hypothesis tests as well were not given credit eg 'if the test statistic is in the critical region the conclusion is to Reject  $H'_0$ 

It was important for candidates to **specifically link** the hypothesis test to its assumption eg the sign test has no distributional assumptions and the Wilcoxon signed rank test with symmetry

for the final marks. Some candidates did not clearly state that the sign tests had no distributional assumptions.

A common incorrect response was to say 'the Wilcoxon signed rank tests is more accurate than the sign test' this did not gain credit as all hypothesis tests are accurate went conducted correctly. The term the specification would require is **powerful** (this was previously mentioned in the exemplar materials from 2206 9ST0 03)

# **Question 5**

The largest question on the exam paper. However, it was clear candidates had been well prepared for ANOVA (this was also true in 2206). Many candidates managed to achieve full marks in (a) or only the final explanation mark. It is important to note that ANOVA is a test on the population mean and therefore it was **expected that the mean was referred to** in both the hypotheses and conclusion.

Where candidates had stated a negative for any of their SS values, they were then unable to get any further method marks.

Candidates should also be encouraged to use the formulas that are given in the formula booklet, there were cases where the student had misremembered the formulas and therefore lost marks.

Students who do this on their calculator should include the total row in their table.

In this series we were generous by allowing candidates to use either test statistic. If candidates are using the tables they must state and compare against **both critical values** (or an interpolated value/calculated value).

Several candidates read ahead and then incorrectly took the data to be a 2-factor table, in these cases they could still gain credit but it is important to note that in an exam candidates may need to read back (as in Q05(e)) in an exam question they will **never be required to read forward** to answer a question.

Part (b) was answered well with many candidates identifying that location could be used as a blocking factor and correctly identifying the design as a Randomised Block Design. In some cases, candidates named another variable as a possible blocking factor, this was not given credit as the question clearly stated 'using this further information' so, as in Q02(d), location was required to be mentioned specifically.

An easy mark was obtained by the vast majority of candidates in (c). It was not required they explain why experimental error had been reduced, though it was pleasing to see many correct attempts at reasons.

Parts (d) and (e) were completed by the vast majority of candidates correctly which was good to see so late on in the paper. Full credit was only given to those testing for a negative correlation as this is the relationship that Joji had described in (d)

# **Question 6**

At this point in the paper some candidates did not make an attempt at this question. Although Cohen's *d* is one of the later topics in the A-level content, there are some easily accessible marks for learnt terminology and methods.

Candidates who were the most successful on this question **broke their answers up** by dealing with the interpretation of the hypothesis test first and then interpreting the value of Cohen's *d* separately.

When interpreting the *p*-value, many students incorrectly interpreted this as the size of the difference between the averages. Rather than identifying that a small *p*-value meant that such an extreme observed outcome would be unlikely under the null hypothesis (quoted from 2206 examiner feedback Q03). For the first mark students were only required to show (or imply) Reject  $H_0$ .

When interpreting the value of Cohen's *d*, many students used incorrect terminology. The specification states the following:

21.2	Know and use Cohen's <i>d</i> in simple situations.	Students should be aware of the standard guideline boundaries for interpreting the value of Cohen's <i>d</i> .
		$0.2 \le d \le 0.5$ small effect size
		$0.5 \le d \le 0.8$ medium effect size
		$0.8 \le d$ large effect size

However, marks were allowed for small or medium as 0.49 was close to the boundaries.

Candidates did not gain full credit if:

- they omitted the word average/mean in the comments (penalised once)
- they did not write two separate comments (one for the interpretation of the *p*-value and one for Cohen's *d*)

# Question 7

This was the most challenging question of the paper; it was encouraging to see many candidates making a good attempt at all parts.

In (a) the first mark was awarded for evidence that the student understood **differences** would need to be found, **context** was required for the second mark as well as identifying that both samples should be used.

Parts (b) and (c) were discriminating for the high achieving students. Similarly, to Q04, those candidates who had invested time into **learning the assumptions** for the hypothesis tests were rewarded.

The paired *t* test was attempted well by candidates, though students should be encouraged to keep going to the end of the paper and **attempt all questions** set as there were some accessible marks.

Also, candidates should understand that a question, and the paper itself, **may not be always be in difficulty order**. For instance, there have been a few occasions in this particular paper where more accessible marks were found later on in a question as well as later on in the paper.

In the final mark of the exam **incorrect or vague terminology was penalised** eg the use of sample/it/data instead of population distribution. It is important that candidates understand that assumptions are about the distribution of the population (or differences for paired tests) and this terminology is expected in their responses as well.

## Summary

Based on their performance on this paper, candidates should be advised to:

- give explanation answers within the context of the question.
- look out for words shown in bold type.
- structure and clearly label answers with multiple parts, as in Q04.
- ensure that they identify key terminology given in the question to ensure answers are detailed enough eg describe..., using this information comment ...,
- avoid the use of the words 'it' and 'data' due to their ambiguity
- multiple attempts at answers will be marked separately, averaged and rounded down. The examiner will not just mark the correct response.

Pearson Education Limited. Registered company number 872828 with its registered office at 80 Strand, London, WC2R 0RL, United Kingdom